IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol.-6, Issue-7, July 2018 E-ISSN: 2347-2693

Big Data and Learning Analytics Model

Sadiq Hussain[1*], Mehmet Akif ÇİFÇİ[2], Josan D. Tamayo[3], Aleeza Safdar[4]

[1]System Administrator, Dibrugarh University, Assam, India
[2] Istanbul Aydin University, Computer Engineering Program, 34295, İstanbul, Turkey
[3]Centro Escolar University Malolos
[4]Bahria University, Islamabad, Pakistan

[*]*Corresponding Author:* sadiq@dibru.ac.in, *Tel.: +91-943524692*

Available online at: www.ijcseonline.org

Accepted: 23/Jul/2018, Published: 31/Jul/2018

Abstract— Big Data opens big opportunities in every corner of the world in almost every companies and industries, viz. banking, stock, agriculture, telecommunications, healthcare and education. With this big opportunity comes with big challenges and issues. Opportunities are increasing as the volume of Big Data is also increasing and predicted to grow enormously because of the technological revolution, which includes but not limited to various mobile devices. The nature of big data using use cases, real-time analysis, data integration, eventually turns big data into a big value. Pressing issues identified in this paper are privacy, processing and analysis and storage. In this paper, we explored various usages of Big Data, methodologies in Big Data and a Learning Analytics Model based on Big Data, as educational entities have sensitive data which are scattered across departments in various formats and need to be processed to gain insight and to make future predictions. Prediction models may be prepared by analyzing the trends from the available historical data. These data models are helpful for data-driven decisions by the authorities.

Keywords— Big Data, Learning Analytics, Hadoop, MapReduce, Big Data Analytics

I. INTRODUCTION

We are living in a digital era where data becomes big, learning becomes deep and the data is increasing exponentially [1]. Artificial Intelligence, Cloud Computing, and Big Data are the talk of the town in the 21st Century. Big Data is the gold mine of this century [2]. In the years to come, I may get a call from a psychological clinic and they may tell me that I am suffering from depression based on the tweets of mine for last one month. I may move with a driverless car and my every move and keystrokes may be tracked. My daughter may inquire about an online course selection system and it would advise her best course suitable for her based on her academic records and some psychometric tests performed online. She may enrol for the best-suited course for her and get instant materials, feedback and guidance from the system and may get some links to understand the topics better. This may be the scenario of the recent future.

From learning analytics to business intelligence, from the health sector to sensor data, the regime of big data continues. McKinsey predicted Big Data as the next frontier for research and innovation[1]. Big data analytics may be categorized as Big predictive, prescriptive and descriptive analytics [3]. The important aspect is to search for pearls in the sea of Big data. Big data is such data that cannot be computed in a single machine of today's world and stored using our traditional database system. Big data is generated from social media, sensor networks, banking transactions, satellite imaging, biomedical projects, business data, genome data etc. By 2020, the number of internet users would be 5 billion with 50 billion devices from the current 2 billion users [4]. The predicted data is 44 times more than the present volume. 90% of data generated in the last two years [5]. In 2012, American administration invested 200 million dollars for research in Big Data [4]. Most of the generated data are semi-structured or unstructured data so the structured databases cannot handle such data. So, the big data comes with various challenges as the volume is huge, data is generated at high speed and for its heterogeneous nature. To understand the rapid growth of unstructured data, one may go through the use of YouTube, Facebook, Twitter, Instagram, Google+ etc. In YouTube, 100 hours of video per minute are uploaded, Facebook users upload 100 TB of data daily, Twitter user publish 175 million tweets daily, 40 million photos per day are uploaded in Instagram while Google+ creates 1 billion accounts per day and the list goes on. One of the biggest challenges is how to accumulate this huge data generation. Laney defined big data with 3V characteristics [1]. They are Volume (data is huge), Velocity (data is coming at a high rate) and Variety (data is of different formats). The characteristics of big data are now characterized by 12V, they are Volume, Velocity, Variety, Veracity, Value, Validity, Viscosity, Visualization, Virility, Volatility, Variability, and Visibility. To process Big Data, Apache Hadoop is a well-established platform. It implements the Google's Map/Reduce computational paradigm that divides the application into pieces and processes each part parallel [4]. So, Hadoop is a powerful programming framework which can process huge data in-parallel on various clusters in the effective fault-tolerant way. In the Map-Reduce framework, Map () and Reduce () are the two main functions. In this distributed computing paradigm of Hadoop, one master module called JobTracker and many slave modules are available called TaskTracker [5]. The Map() function manages the huge data and makes it as key, value pairs in parallel and the key, value are merged by the Reduce() function. From that output, the analytics may be applied to find interesting and valuable information for decision-makers. MapReduce is highly scalable across lots of computing systems and can process zeta bytes of data using batch processing. Hadoop has a variety of components namely HBase, Pig, Hive, HCatalog, Oozie, Zookeeper, Kafka with a paradigm like MapReduce and Hadoop Distributed File System (HDFS) used extensively for Big Data [1]. Hadoop has limitations as well. For efficiency, the Big Data is replicated in multiple locations thus making the Big Data bigger. Hadoop is a complicated system with very limited SQL support. Privacy and security are another concern for Big Data. So, Big Data comes with big challenges and opportunities [2].

In the present scenario of competitive and complex Education System, Big Data has a role to play using Learning Analytics. There is a different Learning Analytics Model based on Big Data. In this paper, we also propose a Learning Analytics Model. The Challenges of the education system can be addressed by proper utilization of Big Data in the educational institutions' concerns.

The rest of the paper is organized as follows: Section 2 presents Literature Review, Section 3 describes Big Data in various fields, Section 4 presents Big Data Methodology and Section 5 describes the Learning Analytics Model and Section VI presents the Conclusion.

II. LITERATURE REVIEW

As different industries have been taking advantages of Big Data since its emergence in the 21st century, such as hospital networks. There's a huge potential value in using Big Data and analytics solutions to overcome challenges and realize the goal of quality, value-based care [6]. Meanwhile, other industries are just beginning to embrace it. Just like in the fashion world, where Big Data is increasingly playing a part in trend forecasting, analyzing consumer behaviour, preference and emotions [7]. Big Data will continue to make a wave in the next generations, therefore companies should be aware that they can benefit from Big Data to gain a competitive advantage, and to make disruptive innovations. Thus, data acquisition, transformation, integration, analysis and visualization can bring business opportunities [8].

If opportunities are laid down to the companies, challenges and issues should also be clear to them so they can make use of Big Data efficiently and effectively. Privacy is identified in [9][10][11][12][13][14] as one of the major concerns in Big Data. [14] stated that anonymous, temporary identification and encryption are the representative technologies for the privacy of data analytics, but the critical factor is how to use, what to use, and why to use the collected data on Big Data analytics. Until regulations on privacy will be more forceful then privacy will remain to be questionable in Big Data. Interestingly [8] wrote that one valuable part of raw data is social media, where citizens publish information relevant to their own situation. Data may be obtained through website crawling or using API´s (Application Programming Interfaces) for social media platforms. It will then be stored, analyzed and results can be sold directly to customers or through brokers/portals. This is a situation where the privacy of citizens using social media is not taking into consideration in data collection.

Another issue in Big Data is the available technology that can process terabytes and petabytes of data and soon to be zettabytes and yottabytes of data. As of this writing, Hadoop is still the leading and widely used platform for processing Big Data. Its core is the Map Reduce, a parallel programming model, inspired by the "Map" and "Reduce" of functional languages, which is suitable for big data processing and analytics functions [15]. Even if this platform changes the economics and the dynamics of large-scale computing and said to be the solution for Big Data [6], there are still areas that needs to be improved such as the generation of multiple copies which makes Big Data enlarged further [11], unavailability of transaction support that makes Hadoop doesn't sound good for frequently changing data [16] and the proliferating amount of data which is making Hadoop inadequate [17].

Attempts for better Big Data solutions are being studied and proposed. [10] identified management of data lifecycle as one of the pressing issues within big data analytics. The huge amount of data coming from different sources produces a gigantic amount at a speed of light, and its usefulness can be revealed when it's still fresh. Thus, data that will be used immediately and data that can be archived for future used must be processed in an instance. To enhance the efficiency of data management, [11] proposed a data-lifecycle that uses the technologies and terminologies of Big Data. A very interesting cycle that can be taken into consideration. Its stages include collection (data collection method includes Log files, Sensing, Mobile equipment, Satellite, Laboratory and Supercomputer), filtering (data can be classified as structure/unstructured, data cleaning/integration), analysis (selection of tools such as, data mining algorithm, cluster, correlation, statistical, regression, legacy codes and indexing), storing (data can be handled through Simple DB, Big Table, Hadoop, MapReduce, Memcache DB and Voldemort), publication (taking consideration of ethical and legal specification, organization and documentation and representation), retrieval and discovery (decision making).

Another issue in Big Data is the analysis of data. Variety of data makes it even more difficult to analyze Big Data. [18] proposed clustering as a solution for the slow speed of analysis of data. Clustering is a data mining tool which finds clusters containing similar data sets. The idea of clustering techniques is to divide data points and group similar data points together and dissimilar data points together. One type of clustering is partitioning-based clustering methods like K-means clustering decomposes dataset into a set of disjoint clusters. Each partition constructed in a dataset represents a cluster and relocates instances by moving them from one cluster to another, starting from an initial partitioning. Another is hierarchical clustering involves creating clusters that have a predetermined ordering, either from top to bottom (Divisive) or bottom to top (Agglomerative). And lastly, there are clustering techniques based on density, the basic idea for the algorithm is that the data space is partitioned into a number of non-overlapping regions or cells, and cells containing a relatively large number of objects are potential cluster centres [19].

It is no doubt Big Data is now the business, there are many opportunities that opened because of this big thing. Other business sectors may consider data mining in their operations, they can start investing and benefit from it viz., fashion industry, agriculture, education and government. Big data can be an overwhelming thing but we must always take into consideration privacy as one of its challenging issues. In any data management cycle, privacy is always an important part of it.

When it comes to platform, Hadoop can handle Big Data and clustering can analyze an enormous amount of big data within a reasonable time. But with the pressing issues of volume, velocity and variety of data to be analyzed promptly, new techniques and methods in data analysis need further exploration and study.

Chatti et. al. [20] designed a reference model based on four dimensions of learning analytics which may provide future research direction in this field. The four dimensions are what, who, why and how in the reference model. Here what means the context, data and environment, who deal with the stakeholders, why considers the objectives and the how dimension means the methods used in learning analytics. The authors also explore various opportunities, challenges and research direction in the area of learning analytics for each dimension.

Picciano [21] explored the concept of big data and learning analytics with special reference to American Higher Education. The author pointed out that although big data and applications related to instructional methods are in infancy, their impact cannot be ruled out in the future. Higher Education administrators may be benefited to make strategies for the students based on the big data with learning analytics. The author provided definitions, concepts, applications and concern about the usage of big data in the field of educational settings.

III. BIG DATA IN VARIOUS FIELDS

Big data have a pivotal role in various fields [22]. Big Data as a term is used for voluminous datasets that have a colossal amount of complex and varied data which is difficult to analyze, visualize or process and result. Nowadays almost all field such as companies, organizations, and enterprises are enjoying having the benefits of Big Data [23]. Furthermore, Big Data is vital for banking, chemistry, finance, healthcare stocks details of which will be presented in this section.

3.1) Big Data in Banking: Because of limited information access by marketing and product development departments in the banks, customer needs could be analyzed with conventional methods such as tracking and segmentation of account transactions [24]. However, thanks to the Big Data technology, banks can now make a lifestyle score model from a mobile phone application, regardless of whether they are customers or not, by monitoring their daily activities without breaching the customers' privacy. By means of Big Data, they develop a new product or service to increase existing customer satisfaction and analyze how non-customers can become potential customers. Customer satisfaction, as well as customer insecurities, can now be analyzed more quickly by dint of Big Data [25]. The complaints about services and the comments on the banks can be processed by banks for better solutions to make their customers happy. Even if banks are not yet able to produce a real-time solution to their customers ' complaints, demands, and suggestions, it will soon be possible with the emerging Big Data technologies. Thus, when the customers enter complaints [26], they will be able to offer any specific solutions to the bank's needs.

Fraud Detection: One of the most important issues in financial sectors is security. There is always a threat of fraud for every bank. That is why the banks are investing heavily in developing fraud detection systems to detect and prevent any scam and to constantly improve their security system by using Big Data [27]. In order to be able to detect fraud in real time, it is necessary to make inferences by processing real-time Big Data that comes from various channels at high speed. At present, it is not possible to detect fraud with traditional systems. Therefore, the banking systems are using Big Data technologies, social media, customer databases and data providers to get instant actions, they are using Big Data for anomaly detections in the banking system [28]. In the use of Big Data technologies and analytics for fraud prevention, it is suggested for analyzing the data firstly in the banks and to remove the activity patterns that are not peculiar to the customers. Information such as the user's key dynamics which channel a user is logged into and from which bank and in which location, is the primary method used to detect fraud [29]. During the extraction of the activity patterns, banking channels should not be considered separately, but all transactions made from all channels must be taken into account in order to catch cross-channel fraud, which is hard to detect.

Big data also allow behavioural authentication. Separation of fraudulent activities from normal activities in the detection of fraud can also be benefited from sharing among mobile and social networks [30]. For example, a bank's fraud monitoring system prevents the generation of false positive rates when the customers are on their vacations. The fraud detection systems recognize the mobile data and with a true positive rate. However, one issue to consider is that it remains within the legal limits in accordance with the legal regulations on customer privacy.

3.2) Big Data in Stock: Big data is used to forecast how the stock of a certain company may change at a certain time. It is benefitted in some ways which are mentioned below. In stock, the first method is 'prediction'[31]. We are benefiting from any technical indicators to assist us to estimate how a stock fluctuates in the next period. Thus, the idea is to use Big Data for the prediction by using machine learning techniques. There are 'features' and 'independent variables' which are named indicators mentioned before. With the help of supercomputers, parallel computing power, hundreds of features can be fed easily with Big Data for training to predict the fluctuate in stock [32]. The second method is 'optimization' [33]. We are using Big Data for optimization problems such as trade execution, portfolio optimization etc. Big data is used with the help of reinforcement learning to optimize stock problems.

3.3) Big data in Agriculture: The history of using smart technology to increase efficiency in agriculture took place in the 70-year period in which we developed the closest, and perhaps the greatest, effects to the dynasty up to hundreds of years ago [34]. The genetic research carried out in the 1980s and the machine follow-up technologies that emerged in the 2000s also influenced the yields from the fields [35]. As a result of all this historical process, agricultural production is realized in today's world with a total value of 3.5 trillion US dollars. Today the whole world is on the verge of a new transformation to create sustainable growth in agricultural production as well as to minimize the possible negative effects of agricultural activity on the natural environment.

In the case of digital technology, which will increase productivity in agriculture, it is necessary to start from the very foundation of agriculture just as it is in all other matters related to agriculture. For this reason, the plant itself is at the zero point of the design of the systems that will provide digital conversion in agriculture. The way to increase efficiency is to know the plant very well, this can only succeed via Big Data.

With the help of Big data, we can tell the farmers the locations of their fields and when to crop them, and what the basic needs are. As seen in the real world the firms have taken Big Data into action to have better results. For example, some biotechnology firms are using sensor data to optimize crop efficiency. They plant test crops and runs simulations to measure how plants react to various changes in condition. They are running tests to find the best environment which constantly changes. Moreover, big data environment constantly regulates to such values as soil composition, temperature, crop growth, water levels of each plant in the test bed.

3.4) Big Data in Enterprise: One can name this era as Enterprise Era when a great increase in demand for various online enterprise applications are taken into consideration [36]. For enterprises around the world, Big Data analytics offers a great many advantages thanks to which business people make correct decisions [37] in real time with fewer expenses when compared with traditional tools. The best example is that Walmart tracks about 1 million transactions per hour [38].

3.5) Big Data in Economy: There is a great amount of data generated. It can be described as an explosion which can be seen as the direct consequence of noteworthy advances in technology. The conventional method was of no use; so with Big Data a new era in the economy has emerged when looked into the companies in the USA it can be seen easily. I.e. when looked at the rise of America's industrial economy for a few decades, it's clear that economic growth has been accompanied by the rise of institutions that provide Big Data [39].

3.6) Big Data in Telecom: As used in many other fields Big data is vital for telecom. There is a competition in the world of telecom services. This is because of gaining more subscribers. In order to have the most subscribers, Big Data is in help because operators are in the belief that big data will play a crucial role in improving their quality of services (QoS) and help them meet their business objectives [40].

Operators meet an uphill challenge when it comes to delivering new, compelling, revenue-generating services which overload the network and cost more than expected. They need big data usage and analytics which can help them make correct decisions by taking a customer, network context and other important aspects of their businesses into consideration. These decisions should be made in real time. Real-time predictive analytics can help leverage the data that resides in their multitude systems, make it immediately accessible and help correlate that data to generate insight that can help them drive their business forward [41].

3.7) Big Data in Healthcare: There are quantity, diversity and speed parameters in the Big Data concept. Data is now being generated in a way that is not easy to manage with traditional data management tools and methods in the health field. In addition, the diversity of produced data and the production speed are also quite high [42]. Data quality in the health system is also an important concern and it is critical that the analysis results are accurate.

Traditionally, the healthcare industry has lagged behind other industries in the use of big data, this is because they do not want to trust protocol [43] based data yet since 2012 this has begun to change. Clinical decision support systems and computerized instruction entry systems within the health and wellness dataset include clinical data, patient data from electronic storage systems, case reports are now taking benefit of Big Data.

Because of the traditional systems medical data - paper files, x-ray films, and notes - were stuck though thanks to Big Data analytics this is a matter of past now. Additionally, routine follow-up blood glucose monitoring, blood pressure evaluations, or regular follow-up data all this kind of data has been managed easily. Over time, health data will be continuously generated, stored, and eventually processed.

In the health system, Big Data sources can be internal and external. They are usually in different formats and are located in different locations as shown in Table 1.

Table 1. Examples of data types and resources

1. Web and social media: interaction with resources such as Facebook, Twitter, LinkedIn, blogs. It may contain websites and smartphone applications for healthcare institutions.
2. Machine-based data: data from sensors, measuring instruments etc.
3. Large transaction data: Health provision requests and other billing.
4. Biometric data: Fingerprint, genetic, handwriting, retinal scan, x-ray and other medical images, blood pressure, heart rate, pulse oximeter readings, etc.
5. Human production data: Electronic data such as medical records, physician notes, and paper documents.

Progress in data management, particularly virtualization and cloud technologies, facilitates the development of platforms for the efficient collection, storage, and use of large volumes of data. In addition, complicated analytical techniques are developed in accordance with the quantity, speed, and diversity of the health data as it develops. Uses of Big Data analytics in health as in Table 2.

Table 2. Various categories of tactical analysis skills.

1. Clinical rules derived from evidence-based studies: Incorporated in some electronic health record systems, they can contribute to the identification of possible risks, especially for general past events.
2. Statistical algorithms derived from evidence-based studies: They are used to mark potential risks in a patient population. They never change. Whatever the context or environment, they produce the same results for the same data.
3. Machine learning: Provides more accurate risk estimation through continuous learning models fed from Big Data.

Big Data generally correspond to the implementation of machine learning algorithms in the analysis of datasets. With machine learning, researchers can investigate possible hypotheses from Big Data sets instead of developing a hypothesis and collecting data from the sample. The finding and matching of verities correlation sets are accomplished through the brute-force classification process combined with the learning and feedback process [44].

Therefore, the results of Big Data machine learning algorithms are perceived as new hypotheses beyond definite estimates. Researchers test limitedly by dividing hypotheses into data sets or re-running algorithms on newly aggregated data sets. If the digital conversion based on information and communication technologies is combined with the use of Big Data, health institutions and organizations can gain real advantages. In early periods when diseases can be treated more easily, diagnoses can be made, individual health status and community health can be managed, and duties in health payments can be determined more straightforwardly and effectively.

The areas in which health care can contribute

1. Clinical procedures: Investigate comparative effectiveness to identify more cost-effective ways of patient diagnosis and treatment.
2. Research and development: Predictive modelling for more affordable, faster and directed research/development processes for drugs and medical devices with less loss, the development of statistical tools and algorithms that will improve the clinical trial design and patient collection, thereby reducing testing errors and accelerating pacemaking of new products.
3. Public health: Analysis of disease patterns to follow public health and rapid response, monitoring of epidemics and contagion, more targeted development of targeted vaccines. The conversion of information to action, which can be used to determine the huge amount of needs, the presentation of services and the prediction and prevention of crises used in public health in general.
4. Evidence-based medicine: Combining and analyzing various structured or unstructured medical, financial, operational, clinical, and genomic data to match treatments with outcomes, identify patients at risk of disease or re-admission, and provide more efficient care.
5. Genomic analytics: Genetic analysis with more

efficient and cost-effective gene sequencing is part of the everyday medical decision-making process.
6. Predictive analysis: Rapid analysis of a large number of provisioning needs to reduce misuse, unnecessary or bad use.
7. Remote monitoring: Receive and analyze data simultaneously and rapidly from devices in a hospital or at home for safety tracking or side effect estimation.
8. Patient profile analytics: Identify patients who can benefit from proactive care or life change by applying advanced analytical techniques such as segmentation and predictive modelling to patient profiling.

In summary, understanding and discovering the relationships and trends between Big Data analytics and data and patterns can make informed decisions, increase the quality of care and reduce costs. A wide range of scenarios can be presented for the potential of Big Data analytics in health. By analyzing patient characteristics, cost and care outcomes, clinically the most cost-effective solutions can be identified and analyzed and tools can be presented.

3.8) Big data in Armed Forces

The standard military operation data by using the platform of the Internet may be of varied use [9]. Military departments are using big data techniques used by businesses to mine databases and collect more information from different data sources- like from terrorist databases, automated cybersecurity systems, drones. Despite helping warfighters on the battlefield, this technology will improve different fields ranging from software development to vehicle maintenance. [45]Organizational, Relationship and Contact Analyzer (ORCA) a piece of software developed by U.S. Military Academy at West Point that attempts to make sense of big data. It is used in intelligence analysis for law enforcement operations against the network of violent street insurgents using algorithmic techniques in social network analysis. This software can determine a set of influential individuals, “degree of membership” for individuals who do not admit to being part of a street gang, and criminal ecosystems by bifurcating gangs into sub-groups. [46] Conditioned-based Maintenance (CBM): The US Air Force officials, data scientists, and other personnel are saving $1.5 million in one year by combining advanced data-handling and analytics tools, both software and hardware solutions, to streamline workflows, increase productivity and efficiencies, and replace scheduled/time-based maintenance track and manage assets, with conditioned-based maintenance (CBM). Specifically, Air Force officials chose to use Teradata’s Aster platform to better manage Aircraft depot data, Inventory and Flight line maintenance. [47] Commanders have to manage and control big data environment comprising of, transactional data, unpredictable pattern of data, historical or point-in-time data, optimized and ad-hoc use of the system and data for inquiry in order to understand and react into real-time tactical situations. With the initiation of unmanned vehicles with sensors, the military has been collecting the huge amount of data at humongous levels. They need dedicated data scientists and innovative software tools to use the extracted information for mission planning. Under a program called Nexus 7, Defense Advanced Research Projects Agency (DARPA) has sent in data scientists and visualizers in Afghanistan. They assisted commanders in solving operational challenges through operating directly with military units.

IV. BIG DATA METHODOLOGY

4.1 Hadoop

Apache Hadoop comes with a distributed file system, on which the storage units of different machines are managed over a network, known as Hadoop Distributed File System (HDFS), that is, however, has file system abstraction and can also work with local file systems [48]. HDFS is a distributed file system designed to store very large files on ordinary hardware/server clusters with streaming data access patterns. Since distributed file systems are network-based, one of the biggest challenges in their management is to truncate node failure without data loss. HDFS manages this problem by distributing and backing up the files it stores in blocks between multiple and different servers (default 3 copies) on the cluster [49].

HDFS is built on a write-once, read-many-times approach. It is therefore ideal for data clusters where analysis operations can be performed many times with different approaches. HDFs are designed to continue to work without a noticeable interruption to the user in the event of failure. Furthermore, Hadoop allows data to be processed in a secure, efficient and scalable way with many features. If an error occurs in a node in another node begins to copy a copy of robust again, thus keeping the data safely. Hadoop works with the principles of parallelism so that the data is processed in parallel, thus preserving its effectiveness. Scalable permits processing of data in the size of petabytes [50]. The HDFS file system records the data across the nodes in the Hadoop cluster. HDFS looks like a traditional hierarchical file system. In other words, files can be created, can be deleted, known file operations can be done. The HDFS architecture consists of special nodes which are:
1. NameNode: Provides metadata service in HDFS.
2. DataNode: Provides storage blocks for HDFS.

The top layer of HDFS is the MapReduce engine consisting of JobTracker and TaskTracker (Figure 1). In HDFS, files are divided into blocks. These blocks are replicated to multiple computers. Block sizes are typically 64 MB. The block size and the number of copies of the data can be determined by the user. All file operations are managed by NameNode. All communications within the HDFS take place via the TCP / IP protocol [51].

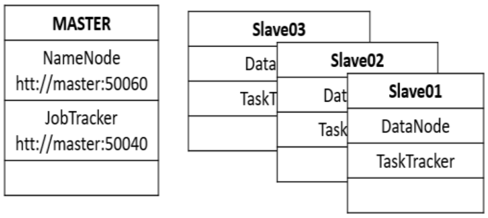

Figure 1. The HDFS architecture

4.2 MapReduce

It is a programming model that allows safe and distributed processing of HDFS data. In the programming model, which is based on Map and Reduce functions. The data blocks in HDFS are filtered by the map function and converted to the desired (key, value) pairs. If the map function is output, the keys are sorted in a group basis and the result is obtained by using the Reduce function. Once the Map and Reduce functions have been written, the corresponding code is executed in the Apache Hadoop environment [52]. Written Map and Reduce code blocks are executed using Apache Hadoop configurations where the corresponding datasets in HDFS are located. The output of the Map function, which runs on the nodes where it is exported, is collected by the TCP / IP network protocol in the distributed system and the results are written in HDFS again [53].

The power of Apache Hadoop comes from the fact that, as mentioned in the previous section, the data is not processed but assigned to the process. In this case, network traffic does not occur in the distributed system, and as the number of nodes in the system increases, the performance increases with the right proportion. There are differences between MapReduce Version 1 and MapReduce Version 2. In MapReduce Version2, there are no JobTracker and Task Tracker concepts. As seen in Fig.2 the YARN resource manager creates an Application Master on any node for every job that is run [54]. Application Master decides which nodes will work on the job by using the Node Manager concepts on each node. In the same way, the Node Managers are also responsible for terminating the work assigned to them [55].

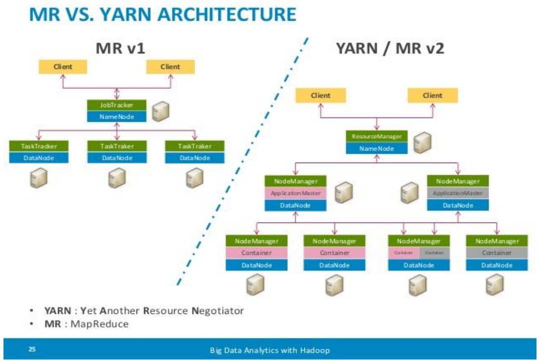

Figure 2. YARN Architecture [56]

V. LEARNING ANALYTICS MODEL

Learning Analytics is not a new research area. It is technology enhanced learning (TEL) and the combination of different areas like pedagogy, computer science, statistics, web science and learning science [57]. It is concerned with different related fields like educational data mining, action analytics, academic analytics and simulates different techniques from machine learning to information retrieval, web dashboards and visualization [58]. There are various existing learning analytics applications used at different educational institutes. Grade Performance System (GPS) is an alert system for the students and used at North Arizona University. Students receive warnings from the system if the students have issues relating to grade, attendance, academic issues.

Purdue University develops a Course Signals System if the student is not performing to the best of his/her ability. For automated tracking and interventions whenever needed of students' progress, Rio Salado Community College developed a learning analytics application called Progress and Course Engagement (PACE) system [59]. Learning is now shifting from knowledge-push to knowledge-pull system [60]. In the knowledge-pull system, a learner is provided knowledge based on learner needs. One of the drawbacks of this system is information overload. So, a good learning analytics model may overcome this by placing a good recommender system in between.

Another area of concern in big learning analytics is that how to integrate and aggregate the student and transaction data from different sources which are also from different formats. So, it is high time to think about Big Education Data as Service (BEDaS). These data-driven models may also be used for decision making by the education policy makers, administrators of education settings and students. According to the New Horizon Report, the Learning Analytics comes fourth among most emerging technologies [61]. We proposed a Learning Analytics model as described below which may be used for courses delivered electronically online. Students may access the data available in the cloud uploaded by various academic institutions. If the data is not available the student may upload their own academic records along with their interests and hobbies to the system. The automatic course selection system has the machine learning logic that helps the students to select from various courses based on the personal data provided by the student. The student selects the course of their choice. The Student would be provided with various online materials, quizzes and assignments. The System analyses the trend of the learning of the student. The system provides instant feedback on the performance of the student. Based on it, the student may be asked for extra sessions and some extra links for the student. The system personalized and adapts new rules based on the individuals' learning needs. So, the system needs big educational data and should be mined efficiently to guide the students based on the machine learning algorithms. So, the system records various transactions of the students course-wise. The previous students' records and analysis may help the new students' learning by using deep learning techniques.

VI. CONCLUSION

Big Data may be considered as the huge numbers of sea-shells. The selected sea-shells may be opened to find the pearls. The Big Data had touched all the spheres including Education and Learning Analytics. Big Data has various opportunities and challenges [11]. Higher Education Institutions data are scattered across departments in various formats. These data need to be processed while guarding the sensitive data to gain insight and to make future predictions. To integrate these unconnected data is a challenge and to overcome it transparency in Institutions is one of the solutions [48]. Prediction models may be prepared by analyzing the trends from the available historical data. These data models are helpful for data-driven decisions by the authorities. Big Data is also helpful in supporting the learner activities and to carry out real-time analysis of learning patterns of the students. Big Data along with machine learning and cloud computing are key players in the future for developing models in Learning Analytics and to use it efficiently by the educational institutions.

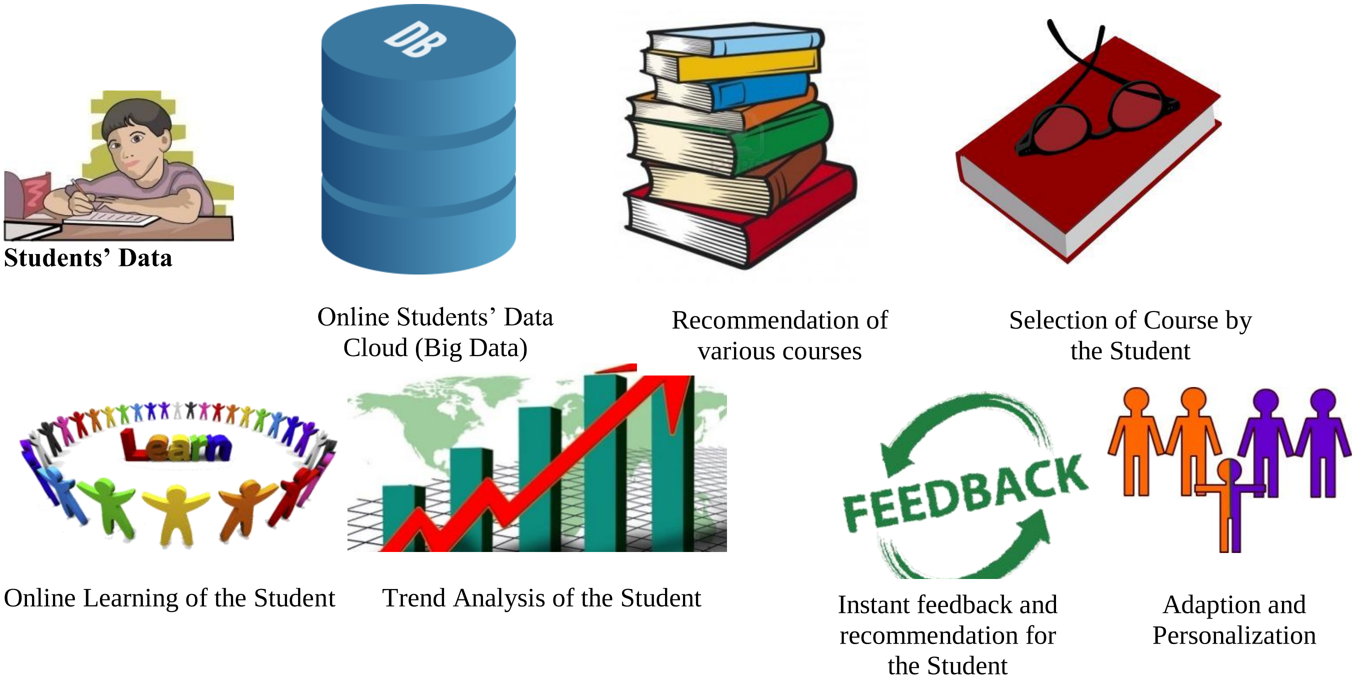
Students' Data
DB
Online Students' Data Cloud (Big Data)
Recommendation of various courses
Selection of Course by the Student
Learn
Online Learning of the Student
Trend Analysis of the Student
FEEDBACK
Instant feedback and recommendation for the Student
Adaption and Personalization

Figure 3. An Overview of the Proposed Learning Analytics System

REFERENCES

[1] Chia-WeiLee, Kuang-YuHsieh, Sun-YuanHsieh and Hung-ChangHsiao, "*A Dynamic Data Placement Strategy for Hadoop in Heterogeneous Environments*", Journal of Big Data Research, vol. 1,pp. 14-22, 2014.

[2] Z. Sun, Zou, H, & K. Strang, "*Big data analytics as a service for business intelligence*", LNCS9373, pp. 200-211, Springer, 2015.

[3] Z. Sun, "*Intelligent Big Data Analytics*", PNG University of Technology, 8 May 2017. BAIS No. 17004, DOI: 10.13140/RG.2.2.32631.83361, 2017.

[4] Divyakant Agrawal, Philip Bernstein, Elisa Bertino, Susan Davidson, Umeshwas Dayal, Michael Franklin, Johannes Gehrke, Laura Haas, Jiawei Han Alon Halevy, H.V. Jagadish, Alexandros Labrinidis, Sam Madden, Yannis Papakon stantinou, Jignesh Patel, Raghu Ramakrishnan, Kenneth Ross, Shahabi Cyrus, Dan Suciu, Shiv Vaithyanathan, Jennifer Widom, "*Challenges and Opportunities with Big Data*", CYBER CENTER TECHNICAL REPORTS, PurdueUniversity, 2011.

[5] D. Laney, "*3D data management: controlling data volume, velocity, and variety*", META Group, Tech. Rep. 2001. [Online]. Available: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf, 2012.

[6] K. Ajit, (n.d.),"*The Emerging Precision, Personalized Medicine and Big Data Analytics Approach in Healthcare: Big Data Analytics in Healthcare"* (Big Data in Healthcare Book 2) Kindle Edition.

[7] K. Al-Barznji, "*Review of big data and big data mining for adding big value to enterprises. Engineering & Education"*, 2, (1), 2017 50 Science, Engineering & Education,2(1), 50-57, 2017.

[8] L. Berntzen., M. Krumova, "Big Data from a Business Perspective. In: Themistocleous M., Morabito V. (eds)

Information Systems”, EMCIS 2017. Lecture Notes in Business Information Processing, vol 299. Springer, Cham, 2017.

[9] H. Bhosale, & D. Gadekar, “*A Review Paper on Big Data and Hadoop*” International Journal of Scientific and Research Publications. , 4(10), 1-7, 2014.

[10] Kevin Taylor-Sakyi, “*Big Data: Understanding Big Data*”, CoRR, abs/1601.04602, 2016.

[11] Nawsher Khan, Yaqoob Ibrar, Hashem I.A.T., Inayat Zakira, Ali, W.K.M., Alam M., Shiraz, M., Gani Abdullah, “*Big Data Survey, Technologies, Opportunities, and Challenges*”, The Scientific World Journal, Volume 2014.

[12] MS Dhruva, “*Survey on Big Data Analytics*,” International Journal of Engineering and Applied Computer Science, vol. 02, no. 06, pp. 181–185, Jul. 2017. http://dx.doi.org/10.24032/ijeacs/0206/02

[13] K. Parimala, G. Rajkumar, A. Ruba, S. Vijayalakshmi, “ Challenges and Opportunities with Big Data”, International Journal Scientific Research in Computer Science and Engineering, Vol 5, Issue 5, pp 16-20, 2017.

[14] Rakesh. S.Shirsath, Vaibhav A.Desale, Amol. Potgantwar, “*Big Data Analytical Architecture for Real-Time Applications*”, IJSRNSC, Volume-5, Issue-4, August 2017.

[15] M. Srinuvasu, A. Koushik, & E. Santhosh, “*Big Data: Challenges and Solutions*”, International Journal of Computer Sciences and Engineering. , 5(10), 2017.

[16] R Gadde,., & N.Vijay, “*A SURVEY ON EVOLUTION OF BIG DATA WITH HADOOP*”, International Journal of Research In Science & Engineering, 3(6), 2017.

[17] M. Memon, S. Soomro, A. Jumani, & M. Kartio,” *Big Data Analytics and Its Applications*”, Annals of Emerging Technologies in Computing (AETiC). ,1(1), 2017.

[18] U. Kazemi, “*Clustering methods in Big data*”, Journal of Embedded Systems and Processing. 2(1,2,3), 2017.

[19] L. Sharma, & K. Ramya, “*A Review on Density-based Clustering Algorithms for Very Large Datasets*”, International Journal of Emerging Technology and Advanced Engineering, 3(12), 2013.

[20] M. A. Chatti, V.Lukarov, H. Thüs, A. Muslim, Yousef, A. M. F., Wahid, U., Greven, C., Chakrabarti, A., Schroeder, U.,” *Learning Analytics: Challenges and Future Research Directions”* eleed, Iss. 10. (urn:nbn:de:0009-5-40350), 2014.

[21] Anthony G. Picciano, “*THE EVOLUTION OF BIG DATA AND LEARNING ANALYTICS IN AMERICAN HIGHER EDUCATION*”, Journal of Asynchronous Learning Networks, Volume 16: Issue 3, 2012.

[22] I.D. Constantiouand, J. Kallinikos, “*New games, new rules: big data and the changing context of strategy*”, Journal of Information Technology, 30(1): p. 44-57, 2015.

[23] T.C. Redman, “*Data quality: the field guide*” Digital press, 2001.

[24] P.Paul, “*Marketing on the Internet*”, Journal of Consumer Marketing, 13(4): p. 27-39, 1996.

[25] S. Mulder, and Z. Yaar, “*The user is always right: A practical guide to creating and using personas for the web*”, New Riders, 2006.

[26] R.L. Villars, C.W. Olofson, and M. Eastwood, “*Big data: What it is and why you should care*”, White Paper, IDC, 2011. 14.

[27] D.D Anderson, “*Debit card fraud detection and control system*”, Google Patents, 1999.

[28] K.V. Rao, and M.A. Ali, “*Survey on Big Data and applications of real-time Big Data analytics*”, Academic Press, 2015.

[29] O. Eisen, “*Methods and apparatus for detecting fraud with time-based computer tags*”, Google Patents, 2015.

[30] A.F. Colladon and E. Remondi, “*Using social network analysis to prevent money laundering*”, Expert Systems with Applications, 67: p. 49-58, 2017.

[31] I.H. Witten et al., “*Data Mining: Practical machine learning tools and techniques*”, Morgan Kaufmann, 2016.

[32] R. Rui, H. Li, and Y.-C. Tu, “*Performance Analysis of Join Algorithms on GPUs*”, Technical Report CSE/14–016, 2014.

[33] A.K ROY, “*Applied Big Data Analytics*”, Paperback. Create Space Independent Publishing Platform, ISBN-10, 2015.

[34] B. Dillon, P. Brummund, and G. Mwabu, “*How complete are labor markets in East Africa?*” Evidence from panel data in four countries. 2017.

[35] M.A. Moisescu, and I.S. Sacala, “*Towards the development of interoperable sensing systems for the future enterprise*”, Journal of Intelligent Manufacturing, 27(1): p. 33-54, 2016.

[36] Zhang, Yingfeng, et al. “*A framework for Big Data-driven product lifecycle management*”, Journal of Cleaner Production, 2017, 159: 229-240, 2017.

[37] A.J. Coale, and E.M. Hoover, “*Population growth and economic development*”, Princeton University Press, 2015.

[38] R. Mahajan, “*Analysis of challenges for management education in India using total interpretive structural modeling*”, Quality Assurance in Education, 24(1): p. 95-122, 2016.

[39] K. Sravanthi, and T.S. Reddy, “*Applications of Big data in Various Fields*, International Journal of Computer Science and Information Technologies (IJCSIT), 6(5): p. 4629-4632, 2015.

[40] A. Thomas, “*Implementing Lean Six Sigma to overcome the production challenges in an aerospace company*”, Production Planning & Control, 27(7-8): p. 591-603, 2016.

[41] A. Collins, and S. Drinkwater, “*Fifty shades of gay: Social and technological change”*, urban deconcentration and niche enterprise. Urban Studies, 54(3): p. 765-785, 2017.

[42] V.K. Jain, “*Big Data and Hadoop*” ,Khanna Publishing, 2017.

[43] V. Tresp, “*Going digital: A survey on digitalization and large-scale data analytics in healthcare*”, Proceedings of the IEEE, 104(11): p. 2180-2206, 2016.

[44] M.J. Mayer, and D.M. Andersen, “*Source-to-processing file conversion in an electronic discovery enterprise system*”, Google Patents, 2017.

[45] B. Akhgar, G.B. Saathoff, H.R Arabnia, R. Hill, et al.: “*Application of Big Data for National Security*”, Elsevier Butterworth-Heinemann, Oxford, 2015.

[46] S. Kulshrestha, “ *Big data in military information & intelligence*” IndraStra Global. doi: 10.6084/m9.figshare.2066640, 2(1), 1–9, 2016.

[47] M.A. Sahin, K. Leblebicioglu, “*Approximating the optimal mapping for weapon-target assignment by fuzzy reasoning*”, Inf. Sci. **255**, 30–44, 2014.

[48] Ben Daniel, “*The Value of Big Data in Higher Education*”, British Journal of Educational Technology, doi: 10.1111/bjet.12230, 2014.

[49] Wadkar, S., M. Siddalingaiah, and J. Venner, *Pro Apache Hadoop*. Apress, 2014.

[50] N. Lohar, “*Content-Based Image Retrieval System over Hadoop Using MapReduce*”, 2016.

[51] L. Alarabi, “*ST-Hadoop: A MapReduce Framework for Big Spatio-temporal Data”* in Proceedings of the 2017 ACM International Conference on Management of Data. ACM, 2017.

[52] P. Zikopoulos, and C. Eaton, “*Understanding big data: Analytics for enterprise-class Hadoop and streaming data*”, McGraw-Hill Osborne Media, 2011.

[53] A.M Hendawi, “*Hobbits: Hadoop and Hive based Internet traffic analysis. in Big Data*”, IEEE International Conference on. IEEE, 2016.

[54] S. Landset, "*A survey of open source tools for machine learning with big data in the Hadoop ecosystem*", Journal of Big Data, 2(1): p. 24, 2015.

[55] I. Polato, "*A comprehensive view of Hadoop research—A systematic literature review*", Journal of Network and Computer Applications, 46: p. 1-25, 2014.

[56] A.P. Kulkarni, and M. Khandelwal, "*Survey on Hadoop and Introduction to YARN*", International Journal of Emerging Technology and Advanced Engineering, 4(5): p. 82-87., 2014.

[57] M.A. Chatti, A.L. Dyckhoff,., U. Schroeder, H. Thus, "*A reference model for learning analytics"*, International Journal of Technology Enhanced Learning 4(5/6), pp. 318-331, 2012.

[58] R. Ferguson, "*Learning Analytics: drivers, developments and challenges*", International Journal of Technology Enhanced Learning 4(5/6), pp. 304-317, 2012.

[59] M. Crush, "*Monitoring the PACE of student learning: Analytics at Rio Salado Community College*", Campus Technology, 2011.

[60] S. Dawson, D. Gasevic, G.Siemens, S. Joksimovic, "*Current State and Future Trends: a citation network analysis of the learning analytics field*", Proceedings of the Fourth International Conference on Learning Analytics & Knowledge, ACM New York, NY, USA, pp. 231-240, 2014.

[61] L. Johnson, S. Adams, and M. Cummins, "*The NMC Horizon Report: 2012 Higher Education Edition"*, Austin Texas: The New Media Consortium, 2012.

Authors Profile

Sadiq Hussain is System Administrator at Dibrugarh University, Assam, India. He received his PhD degree from Dibrugarh University, India. His research interest includes data mining and machine learning. He is associated with Computerization Examination System and Management Information System of Dibrugarh University.

Mehmet Akif CIFCI received the BSc degree in the English language from the University of 9 September, Izmir, Turkey and the MSc in artificial Data mining from Istanbul Aydin University, Istanbul, Turkey. He is now working toward the PhD degree in computer science in Istanbul, Turkey. His research focuses on applied machine learning, Networks science and mining, deep learning.

Josan D. Tamayo is the Program Head of the Computer Education Department at Centro Escolar University Malolos. She finished her Bachelor of Science in Information and Computer Science at the University of the Cordilleras (formerly Baguio Colleges Foundation) and Master of Science in Information Technology at St. Linus University. She is an IBM Certified Academic Associate in DB2 9 Database and Application Fundamentals. She also acquired the NC II Certification for Hardware Servicing. And currently pursuing her Doctor of Information Technology at the La Consolacion University of the Philippines.

Aleeza Safdar is completing her MS in Software Engineering from Bahria University Islamabad, Pakistan. Her research interest includes Requirements prioritization in Software Development, how technology and psychology can go side by side like patterns of psychology being modelled by technological tools, like we have agent-based modelling tools and most commonly used language i.e Netlogo.